SYMBIOSIS
SCHOOL OF ECONOMICS

॥वसुधैव कुटुम्बकम्॥

STOCKS ANALYSIS AND PREDICTION USING BIG DATA ANALYTICS AND MACHINE LEARNING

Submitted By:
Aisharriya Dasgupta- 23060242003
Vidit Vaswani- 23060242098

Table of Contents

ABSTRACT

Big data analytics has become essential in transforming stock market analysis. This study explores the integration of machine learning (ML) models with big data techniques to predict daily stock returns, focusing specifically on US oil stocks. We develop a robust analytical system utilizing Apache Spark's machine learning library, capable of processing real-time financial data from Yahoo Finance. The system employs various algorithms, including Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM), to analyze time-series data and predict stock prices effectively. To enhance prediction accuracy, the XGBoost algorithm is employed for feature engineering, emphasizing the most relevant data points. Evaluation of results using R-squared metrics indicates that the LSTM model significantly outperforms the ARIMA model in prediction accuracy. This research highlights the critical role of big data in financial forecasting and presents a scalable, efficient approach for real-time stock market analysis, providing valuable insights for investors and analysts in a dynamic financial environment.

INTRODUCTION

The advent of big data has revolutionized various industries, including finance, healthcare, and information technology (IT), by enabling the processing and analysis of vast volumes of data, leading to more insightful and informed decision-making. In the financial sector, particularly, the integration of big data and machine learning (ML) technologies has significantly enhanced the ability to analyze complex market dynamics. Stock market forecasting, historically considered a challenging endeavor due to its inherent volatility and the myriad factors influencing price movements, has witnessed substantial improvements with the advent of big data analytics.

Traditionally, stock price predictions have relied on limited datasets and conventional statistical methods, which often fail to capture the intricate and nuanced patterns that characterize market behavior. However, big data analytics offers a powerful solution by providing the capability to manage and analyze large datasets effectively, leading to more accurate and timely stock market predictions. By harnessing the power of big data, analysts can now consider a wider array of variables, including market trends, economic indicators, and sentiment analysis, ultimately enhancing their forecasting accuracy.

This paper proposes a novel system specifically designed for predicting daily stock gains, with a particular focus on US oil stocks. The proposed system utilizes real-time data obtained from Yahoo Finance, which is processed through advanced big data technologies such as Apache Spark and Hadoop. These technologies enable the system to handle and analyze massive volumes of stock data in real-time, ensuring that the predictions are based on the most current market conditions.

To achieve high accuracy in stock price predictions, the system employs sophisticated machine learning models, including AutoRegressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) networks. ARIMA is a widely used time-series forecasting method that captures linear relationships in data, while LSTM, a type of recurrent neural network, excels at modeling complex, non-linear patterns over time. By integrating these advanced modeling techniques, the system is designed to provide precise daily predictions of stock prices, thereby assisting traders and financial institutions in making informed and timely decisions.

A significant challenge in stock market analysis lies in managing the enormous volume and complexity of stock data. Traditional prediction models often struggle to accommodate the dynamic nature of the market and fail to leverage the full potential of big data. This paper emphasizes the importance of establishing a robust data processing pipeline that can handle both structured and unstructured data in real-time. Such a pipeline not only enhances the accuracy of predictions but also allows for the inclusion of diverse data sources, improving the overall analytical framework.

LITERATURE REVIEW

Several studies have been conducted on the application of machine learning and big data analytics in stock market predictions. Traditional methods for stock prediction, such as time-series models and volatility-based models, often fall short in capturing the complexities of stock price movements. Researchers have explored various alternatives, including data mining techniques and machine learning algorithms, to improve prediction accuracy.

For instance, Zhao and Wang (2015) introduced an outlier mining algorithm to predict stock market trends based on high-frequency data. Their approach demonstrated the potential of using anomaly detection in identifying significant trends that affect stock prices. Similarly, Tiwari et al. (2017) employed data analytics tools like ARIMA and neural networks to forecast stock prices in the Nifty 50 market, showing that machine learning techniques could improve prediction accuracy over traditional methods.

In another study, Singh and Thakral (2017) analyzed stock indexes and constituents to identify high-performing stocks. By leveraging statistical analysis and historical data, they provided a systematic approach to selecting profitable stocks. Their research emphasizes the need for continuous monitoring and analysis to understand stock market trends.

Big data analytics has also been employed to uncover fraud in financial markets. According to Peng (2019), financial institutions use large datasets to identify illegal trading activities and develop high-frequency trading systems. This underscores the importance of using robust data processing frameworks like Hadoop and Apache Spark in financial applications.

The use of machine learning models, such as ARIMA and LSTM, has been extensively explored in the context of time-series forecasting. While ARIMA models are effective in identifying linear patterns in stock data, LSTM models are better suited for capturing long-term dependencies in sequential data, making them ideal for stock price prediction. The application of XGBoost for feature selection further enhances prediction accuracy by eliminating irrelevant variables and focusing on the most significant features.

METHODOLOGY

The system proposed in this paper employs a big data analytics framework, leveraging Apache Spark for real-time data processing. The methodology is divided into several key stages:

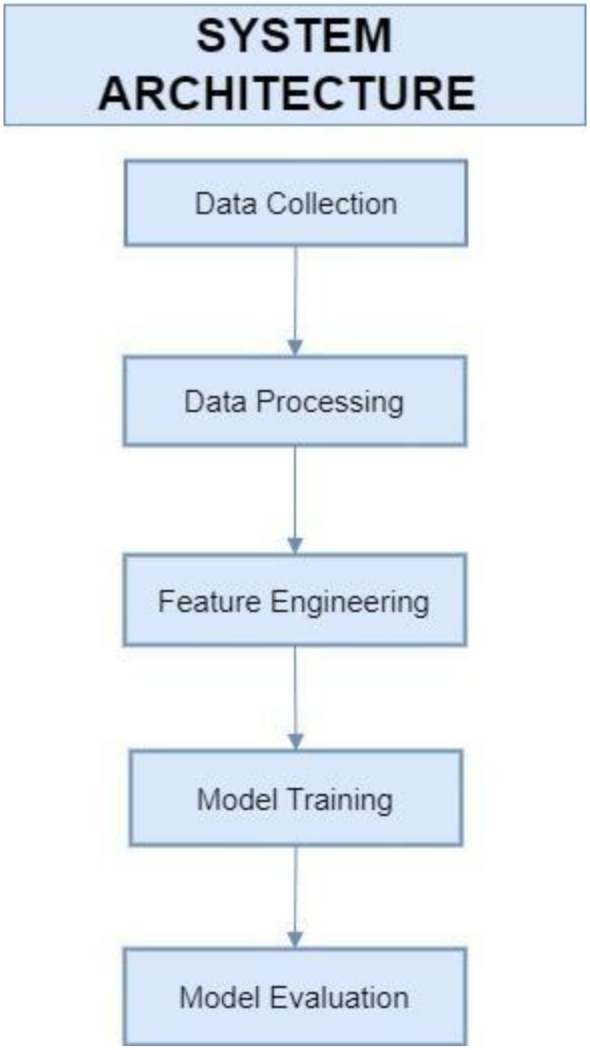
1. Data Collection: Real-time stock data for 13 US oil companies is collected from Yahoo Finance. This data is divided into training and testing datasets to facilitate model development.

2. Data Processing: The stock data is normalized using data scaling techniques to ensure consistency across different features. This step is crucial for improving the performance of machine learning models.

3. Feature Engineering: The XGBoost algorithm is employed for feature selection. XGBoost helps in identifying the most relevant features in the dataset, which significantly improves the prediction accuracy of the machine learning models.

4. Model Training: Two models, ARIMA and LSTM, are used for stock price prediction. ARIMA is applied for time-series analysis, while LSTM is chosen for its ability to capture long-term dependencies in the data.

5. Model Evaluation: The performance of both models is evaluated using the R-squared metric, which measures the accuracy of the predictions. Higher R-squared values indicate better model performance.



Modules

1. Data Reading with PySpark:

- Load stock data from a dataset file using PySpark.
- Initialize Spark and configure a Spark Streaming Context.
- Create a Spark session.
- Read the dataset as either a stream or in batches, utilizing PySpark classes.
- Display the dataset to confirm successful loading.

2. Data Normalization:

- Normalize the values in the dataset to maintain consistency in scale.

- Implement normalization techniques to adjust dataset values.
- Normalization is essential for features with varying units or ranges, allowing algorithms to learn more effectively from the data.

3. Feature Engineering with XGBoost:

- Use the XGBoost algorithm for feature engineering, focusing on selecting relevant features while disregarding irrelevant ones.
- XGBoost enhances the model's predictive accuracy by emphasizing the most significant features.

4. Training the ARIMA Model:

- Train the ARIMA model using the cleaned dataset.
- Specify the order (p, d, q) for the ARIMA model during training.
- ARIMA is a time series model that analyzes data changes over time.

5. Evaluating the ARIMA Model:

- Assess the performance of the ARIMA model.
- Apply the trained model to a test dataset.
- Use the Rsquared metric to evaluate the model's effectiveness in explaining the variance in the test data.

6. Training the LSTM Model:

- Train the LSTM model using the preprocessed dataset.
- Set up a Sequential model in Keras, incorporating LSTM layers.
- Train the LSTM model to capture longterm dependencies in sequential data.

7. Evaluating the LSTM Model:

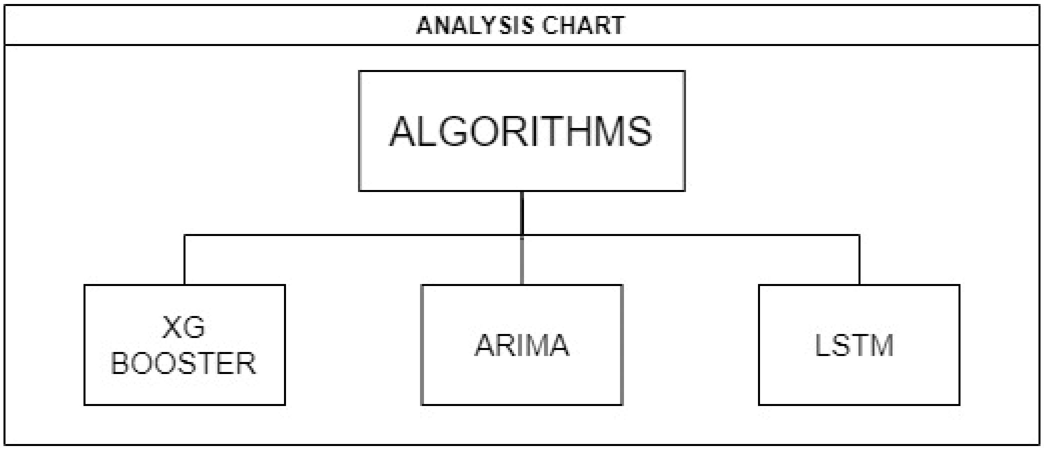
- Assess the performance of the LSTM model.
- Test the trained LSTM model on the dataset.
- Evaluate performance using metrics like Rsquared.
- Compare the Rsquared values of the LSTM and ARIMA models to determine which performs better.

IMPLEMENTATION OF ALGORITHMS

The ARIMA model is a statistical technique used for time-series forecasting. It consists of three main components: AutoRegression (AR), Integrated (I), and Moving Average (MA). In this paper, ARIMA is employed to predict stock price movements based on historical data. ARIMA models are effective in identifying trends and seasonality in the data, which are crucial for stock price prediction. However, ARIMA models are limited in their ability to capture non-linear patterns in the data.

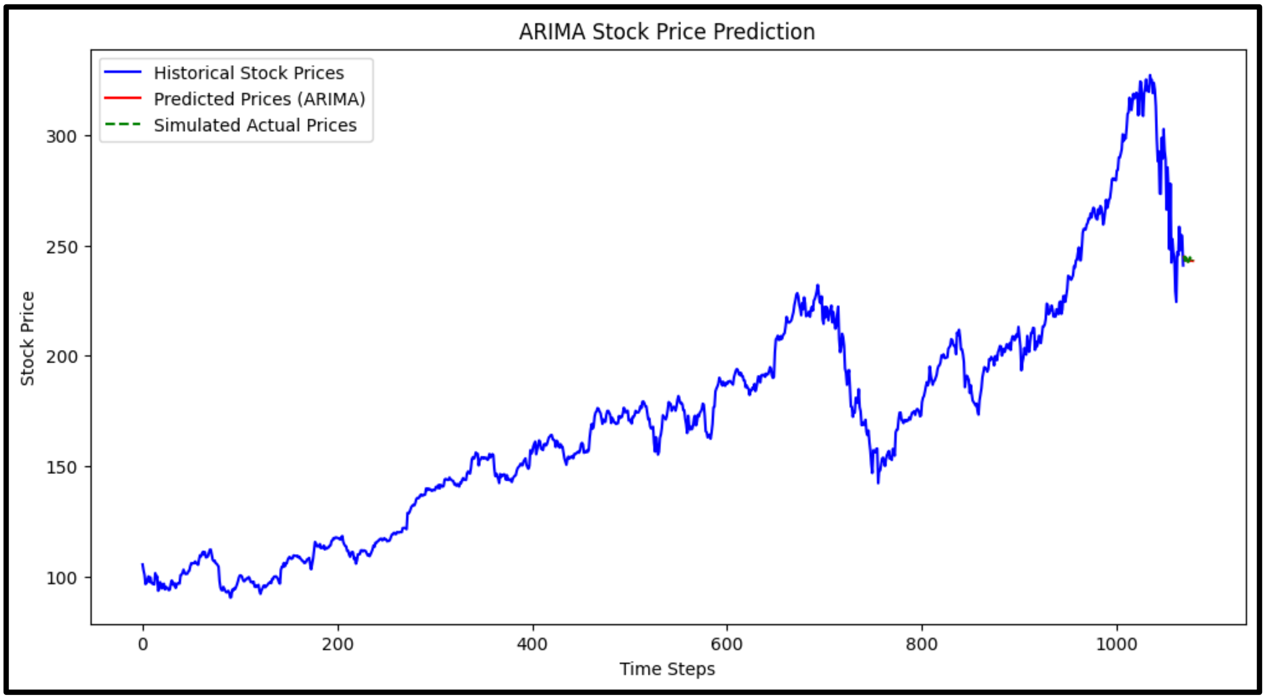
In contrast, LSTM is a type of recurrent neural network (RNN) specifically designed to handle long-term dependencies in sequential data. LSTM overcomes the vanishing gradient problem associated with traditional RNNs by maintaining a cell state that can store information over long periods. This makes LSTM particularly well-suited for stock price prediction, where past data can have a significant influence on future trends.

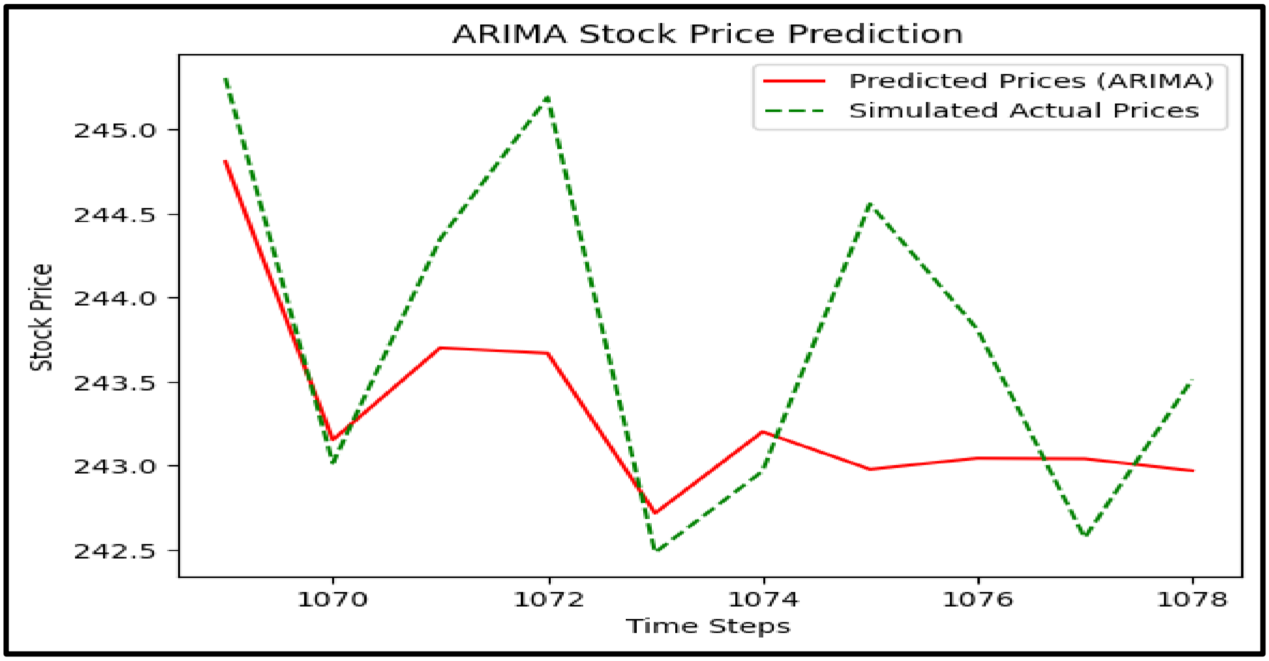
The XGBoost algorithm is used for feature selection in this study. XGBoost is a gradient boosting algorithm that builds multiple decision trees to improve prediction accuracy. By focusing on the most relevant features, XGBoost enhances the performance of both ARIMA and LSTM models.

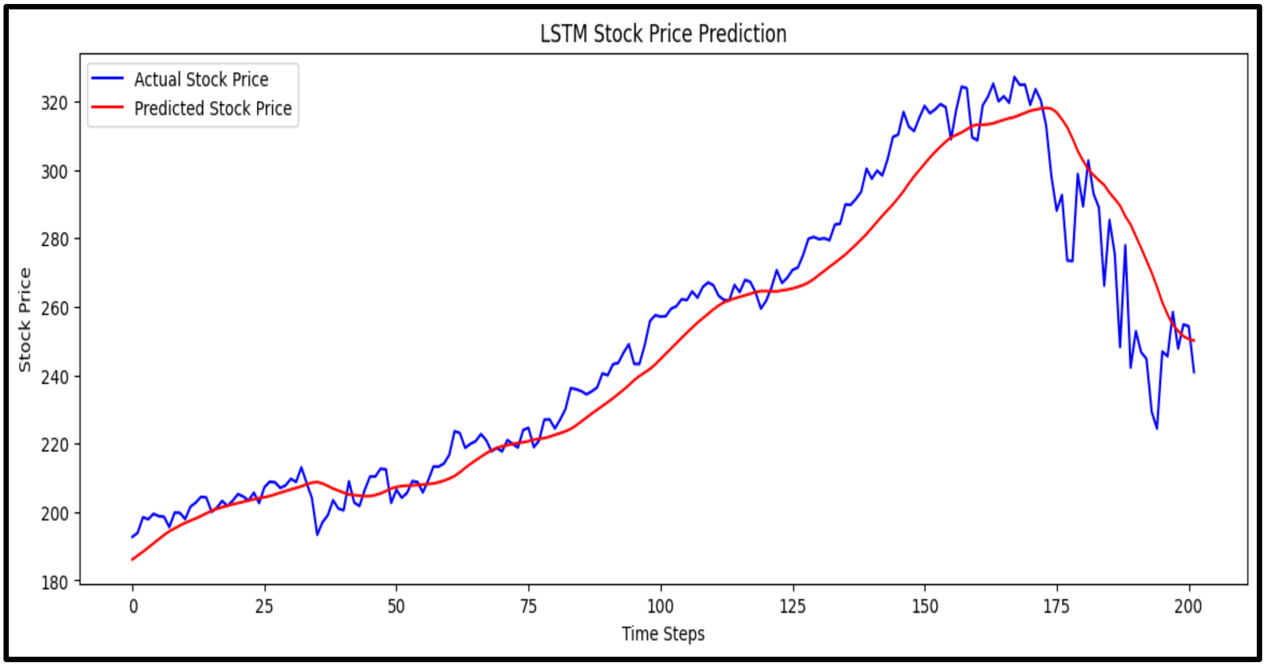


ANALYSIS OF RESULTS

The results of the study show that the LSTM model outperforms the ARIMA model in terms of prediction accuracy. The R-squared value for the LSTM model is 0.91, indicating a high level of accuracy in predicting stock prices. In contrast, the ARIMA model achieves an R-squared value of 0.308, suggesting that it is less effective in capturing the complexities of stock price movements.




ARIMA Stock Price Prediction
Predicted Prices (ARIMA)
Simulated Actual Prices
Stock Price
245.0
244.5
244.0
243.5
243.0
242.5
1070
1072
1074
1076
1078
Time Steps


LSTM Stock Price Prediction
Actual Stock Price
Predicted Stock Price
Stock Price
320
300
280
260
240
220
200
180
0
25
50
75
100
125
150
175
200
Time Steps

DISCUSSIONS

This paper explores the application of big data analytics to facilitate rapid assessments and predictions in the stock market. The stock market is inherently unpredictable, and the inability to accurately forecast stock values can lead to significant financial losses for investors. In light of this challenge, our research aims to develop a methodology for identifying stocks with positive daily return rates, which could potentially enhance trading strategies.The proposed method functions as a Hadoop-based system that leverages historical data to learn patterns and trends. By analyzing this past data, the system can make informed decisions about which U.S. stocks are worth trading, taking into account real-time market fluctuations. This approach not only aims to identify profitable stocks but also seeks to enhance the efficiency of trading operations by providing timely insights into market conditions.

In addition to developing this predictive framework, we recognize the need for continuous improvement and adaptation of our methodologies. As part of our future research initiatives, we plan to implement scheduling tools to automate various analytical processes. Automation will streamline the analysis workflow, enabling us to provide regular updates and recommendations on U.S. stock trading strategies. Furthermore, we aim to enhance our predictive capabilities by exploring advanced machine learning techniques, specifically neural network models. Unlike traditional linear regression models, which may struggle to capture complex relationships within the data, neural networks can learn intricate patterns and dependencies in the dataset. By incorporating this approach, we hope to achieve more robust and accurate predictions regarding U.S. stock prices, ultimately leading to better-informed trading decisions.

CONCLUSION

The study concludes that integrating big data analytics with machine learning significantly enhances the accuracy of stock price predictions. The use of advanced algorithms like LSTM allows for better handling of complex patterns in financial time-series data. The authors recommend further exploration into hybrid models that combine the strengths of different machine learning techniques to improve prediction outcomes. They suggest that future research could focus on incorporating sentiment analysis from social media and news sources to enrich the predictive capabilities of their models.

The findings have practical implications for investors and financial analysts, as they underscore the potential for big data analytics to transform stock market forecasting. By leveraging these technologies, stakeholders can make more informed decisions, ultimately leading to better investment strategies. In summary, this paper presents a compelling case for the integration of big data analytics and machine learning in stock price prediction, demonstrating significant advancements over traditional methods.

Appendix

Google Drive link for the code and dataset used for the study:
https://drive.google.com/drive/folders/1ryKpC22Cz0N3apJzC8dlMEl1ihMoSLs6?usp=sharing

References

[1] L. Zhao and L. Wang, “Price Trend Prediction of Stock Market Using Outlier Data Mining Algorithm,” in
2015 IEEE Fifth International Conference on Big Data and Cloud Computing, Dalian, China, 2015, pp. 93–98.
[2] M.D. Jaweed and J. Jebathangam, “Analysis of stock market by using Big Data Processing Environment”
in International Journal of Pure and Applied Mathematics, Volume 119
[3] S. Tiwari, A. Bharadwaj, and S. Gupta, “Stock price prediction using data analytics,” in 2017 International
Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, 2017, pp. 1–5
[4] Kumar, B.S. et al. (2024) STOCKS ANALYSIS AND PREDICTION USING BIG DATA ANALYTICS AND MACHINE LEARNING. Available at: https://www.ijcrt.org/papers/IJCRT24A4474.pdf (Accessed: 03 October 2024).